Coarse-Grained Density Functional Theory Predictions via Deep Kernel Learning

Ganesh Sivaraman[†] and Nicholas E. Jackson[*,‡]

[†]Data Science and Learning Division, Argonne National Laboratory, Lemont, IL 60439, USA

[‡]Department of Chemistry, University of Illinois at Urbana-Champaign, 505 S Mathews Avenue, Urbana, Illinois, 61801, USA

E-mail: jacksonn@illinois.edu

Abstract

Scalable electronic predictions are critical for soft materials design. Recently, the Electronic Coarse-Graining (ECG) method was introduced to renormalize all-atom quantum chemical (QC) predictions to coarse-grained (CG) molecular representations using deep neural networks (DNN). While DNN can learn complex representations that prove challenging for traditional kernel-based methods, they are susceptible to overfitting and the overconfidence of uncertainty estimations. Here, we develop ECG within the GPU-accelerated Deep Kernel Learning (DKL) framework to enable CG QC predictions of a conjugated oligomer using range-separated hybrid density functional theory. DKL-ECG provides accurate reproduction of QC electronic properties in conjunction with prediction uncertainties that facilitate efficient training over multiple temperature data sets via active learning. We show that while active learning algorithms enable efficient sampling of a more diverse configurational space relative to random sampling, the predictive accuracy of DKL-ECG models is effectively identical across all active learning methodologies employed. We attribute this result to the low conformational barriers of our test molecule and the redundant sampling of configurations induced by Boltzmann sampling, even for distinct temperature ensembles.

Introduction

The accurate and scalable modeling of electronic structure in non-crystalline morphologies underscores a broad array of critical soft materials applications.[1–4] To perform scalable *in silico* characterizations of morphology-dependent electronic properties, one typically relies upon the combination of all-atom molecular dynamics (MD) simulations and density functional theory (DFT) electronic structure calculations. All-atom MD samples the configurational distribution function at nanometer scales, with configurational snapshots being extracted and used as input for DFT calculations of energies and couplings.[4–7] These calculations provide a means of parameterizing model Hamiltonians, reaction rates, and transport models useful for understanding electronic properties of heterogeneous soft materials. While this computational paradigm has seen usage, its poor scaling and computational redundancy[8] for systems beyond nanometer spatiotemporal scales strongly limits its usage for the *in silico* design of electronic functionality in soft materials.

For molecular and polymeric systems exhibiting slow relaxation dynamics or mesoscopic ordering, coarse-grained (CG) models supersede all-atom models for configurational sampling.[8–10] Provided the need for all-atom MD to keep track of atomic degrees of freedom, the accessible spatiotemporal scales are intrinsically limited ($\sim$10's nm and $\sim$100's ns). Contrastingly, CG simulations utilize pseudoatoms that are averages over collections of atomic degrees of freedom, facilitating computational acceleration. Moreover, the renormalized intermolecular interactions between CG pseudoatoms are generally softer, accelerating system dynamics and improving configurational sampling. These reduced models can also enhance conceptual understanding when analyzing the results of large-scale simulations. To maintain rigorous connections to the statistical mechanics of underlying all-atom models, a variety of frameworks have emerged with important applications in both biological and materials systems.[11–14] Such CG approaches have also emerged in the context of machine learning (ML) driven dimensionality reduction of configurational space,[15–19] further emphasizing the important nature of CG modeling across molecular domains.

Despite these benefits, CG modeling places obvious limitations on the ability to solve the Schrodinger equation for the underlying atomic degrees of freedom. By virtue of CG models averaging over atomic degrees of freedom, the ability to deduce electronic structure directly from CG configurations is lost, as one cannot solve the Schrodinger equation for CG pseudo atoms. This fundamental issue has led to a variety of *ad hoc* computational protocols in which CG simulations are performed, all-atom representations are backmapped[20–22] onto CG configurations, resampled, and subsequent quantum-chemical (QC) calculations are performed.[5,7] This approach exhibits high computational cost, is often defined in an *ad hoc* manner, and is strongly limited by the poor scaling of QC calculations. Methods capable of bypassing these complicated protocols and performing electronic structure calculations directly at the CG model resolution exhibit potential for rapidly accelerating both materials understanding and design efforts across disciplines.

Recently, we have introduced a systematic computational strategy for CG electronic predictions known as Electronic Coarse-Graining (ECG).[23,24] Due to the theoretical intractability of performing systematic renormalizations over both electronic and nuclear degrees of freedom, machine learning (ML) approaches have been employed. Previous efforts have used deep neural networks (DNN) with ECG to provide semi-empirical QC predictions of molecular orbital energies, optical properties, and charge densities of conjugated oligomers.[23,24] While existing ECG approaches exhibit considerable promise, within their current formulation they (i) are unable to provide accurate predictions of model uncertainty and (ii) suffer from the need to develop an extensive training set of atomic configurations and their associated electronic properties. To improve ECG models, concerted efforts must be made to both incorporate prediction uncertainties as well as reduce the required amounts of training data, especially when learning electronic properties associated with high computational cost QC methods. For organic materials, range-separated hybrid DFT is a common and reliable QC methodology,[25] representing a desirable target for ECG predictions extending beyond semi-empirical methodologies.

Gaussian process regression (GPR) represents a class of Bayesian supervised learning models capable of providing prediction uncertainties using only learning a small number of kernel parameters.[26] GPR can also provide high predictive performance under low data limits compared to DNN. An example can be drawn from two recent studies on learning the potential energy surface of molten NaCl where the GPR-based Gaussian approximation potential (GAP) was fitted with $\sim$ 1000 training samples computed with DFT whereas a comparable DNN based approach required 100 times more samples to reproduce the correct structural properties for molten NaCl.[27,28] GPR has found applications such as fitting the potential energy surface for atomistic simulations,[29–31] transition state searches,[32] CG simulations,[33,34] and materials design.[16,35] Despite GPR's utility, GPR kernels are often unable to learn complex representations of high dimensional feature inputs such as images. DNN's on the other hand have shown considerable success in learning complex feature representations resulting in impressive predictive performance,[36] though can be susceptible to overconfidence in the estimation of uncertainties.[37] Here, we combine *"the best of both worlds"* by leveraging the complex representation learning power of DNN with the Bayesian nature of GPR through a Deep Kernel Learning (DKL)[38] framework. In DKL, the DNN processes the high dimensional input feature and generates an intermediate representation, which is then fed into the GPR kernel. We have conceived such a hybrid strategy with two goals: (1) provide ECG prediction uncertainties and (2) enable the use of data poor ECG models via AL query strategies that decrease the amount of required training data and facilitate the use of high computational cost QC methods.

Active learning (AL) has emerged as a powerful paradigm in which an algorithm judiciously selects data points for labeling and inclusion in training data that optimally improve model performance. Here, labeling would mean an expensive evaluation involving QC calculations.[39] AL has been employed to facilitate dramatically improved computational efficiency in the DFT screening of oxidation potentials in redox active molecules,[40] as well as improved ML force-fields,[41–45,45,46] including molten salts.[47,48] Drug design has combined AL with ML regression models to great effect in identifying new candidate compounds.[49] Recent advances in AL for regression show exceptional promise in bypassing the canonical AL query strategies.[50,51] For ECG, AL represents a potential means of extracting distinct conformations from the training data sets for QC labeling at minimal computational cost.

In this work, we cast ECG in a format that combine DNN's with GPR though a DKL framework with the goal of arriving at a range-separated hybrid DFT accurate ECG parameterization with associated prediction uncertainties. First, we describe the details of data set generation via MD, DFT and CG mapping operators. Then we detail the DKL regression framework utilized for predictions that facilitates the incorporation of uncertainties. This is followed by the description and implementation of three sampling strategies: random query (RQ), uncertainty query (UQ), and expected model output changes (EMOC). We then apply the AL DKL models to learn the electronic structure of a conjugated oligomer across multiple simulation temperatures. We then conclude and summarize the takeaways of this work.

Methods

Data Generation

We utilize a hexamer of poly(3-hexyl)thiophene with alkylic side chains cleaved to methyls, denoted sexi(3-methyl)thiophene (S3MT), employed in previous work.[23] S3MT is used as a molecule representative of thiophene incorporating chemistries found in a broad array of organic semiconductor applications.[52] The soft bond, angle, and dihedral degrees of freedom of S3MT coupled with strong π-electron conjugation induce substantial configurational dependence of the electronic structure.

To generate representative training sets of atomic S3MT configurations, all-atom MD simulations of a single S3MT molecule are performed using the optimized potentials for liquid simulations (OPLS) force field[53] with partial charges and intermonomer dihedral potentials obtained via QC parameterization in previous works.[23] MD simulations are performed under two sets of constraints (rigid and flexible) and four temperatures (50K, 300K, 600K, 1000K). "rigid" constraints treat each thiophene monomer along the backbone as a rigid monomer and "flexible" corresponds to a standard MD simulation without constraints - further details may be found in previous work.[23] Rigid constraints applied to conjugated molecules is a standard protocol for treating charge transport in disordered organic semiconductors,[54] but naturally will limit the configurational fluctuations to be only those between thiophene monomers. By comparing "rigid" and "flexible" configurations, one can separate the role that temperature plays in improving sampling of dihedral degrees of freedom from the full configurational distribution function.

Temperature is controlled using a Langevin thermostat with a damping parameter of 100 fs^{-1}. Lennard Jones and Coulomb interaction use short-range cutoffs of 25.0 Angstroms and shrink-wrapped boundaries; these parameters ensure that the single molecule only interacts with itself and no images. Flexible and rigid simulations use timesteps of 1 fs and 5 fs, respectively. Flexible MD simulations consist of a 100 ps heating from 300K to 1000K, 100 ps sampling at 1000K, 1 ns annealing from 1000K to the target temperature T, and 40 ns simulation at T from which configurations are drawn at 4 ps intervals. Rigid MD simulations consist of 500 ps heating from 300K to 1000K, 500 ps sampling at 1000K, 5 ns annealing from 1000K to the target temperature T, and 200 ns simulation at T from which configurations are drawn at 20 ps intervals. Finally, we generated held out test sets for model validation by performing dihedral restrained all-atom molecular dynamics simulations of S3MT under both rigid and flexible constraints. In these simulations, the minima of the harmonically restrained dihedral potentials ($k = 5kcal/mol$) are selected via latin hypercube sampling

(LHS).[55] 100 distinct sets of dihedral restraints are applied. For each set of restraints, S3MT is equilibrated for 50 ps and sampled every 4 ps over 200 ps at 300K, generating 20 distinct configurations per set of dihedral constraints - this leads to rigid and flexible test sets consisting of 2000 configurations each. These configurations are entirely separate from the Boltzmann sampled configurations, but span a broad range of S3MT conformations, presenting an orthogonal test set for ECG model validation. All simulations are performed using LAMMPS.[56]

S3MT configurations derived from MD simulations are input to $\omega B97X{-}D3/def2{-}SVP$ calculations using Orca[57] to "label" the electronic structure of all S3MT configurations. In this work we focus on the highest occupied molecular orbital (HOMO) energy as the "label" for specific S3MT configurations due to its relevance for hole transport. As we are interested in developing electronic prediction methods that operate using only CG representations of chemistries, we select a CG mapping operator that defines CG coordinates as linear combinations of all-atom coordinates. We use the systematic graph-based CG algorithm of Webb[58] to generate a hierarchy of CG mapping operators from which we select a 3-bead per thiophene monomer representation, the mapping of which is provided in the SI. The 3-bead per thiophene mapping is optimal for being able to uniquely define all relative 3-dimensional orientations between thiophene monomers within a rigid body approximation.[59] This mapping is applied to every all-atom configuration to create CG configurations. For input into DKL, these CG configurations are then featurized using the distance matrix between all CG beads, from which the upper triangle of the CG distance matrix is flattened and used as an input feature to DKL.

Deep Kernel Learning

To facilitate the integration of AL algorithms with ECG, prediction uncertainties. Here we select GPR as a systematic framework for generating both prediction means and uncertainties. However, it is well-known that the performance of GPR depends strongly upon, and can be limited by, the specification of the kernel function. In the context of this work, we show a representative example of the limitations of traditional kernel-based GPR in an explicit numerical ECG experiment available in Figure S1.

To combine the powerful representational flexibility of DNN, as demonstrated in our previous ECG work,[23,24] with the GPR framework we employ the concept of DKL introduced by Wilson et al.[38] In this approach, DNN learn a flexible representation of the input feature, which is then used as input into a non-parametric GPR kernel, providing scalable closed form covariance kernels for GPR. DKL utilizes a base kernel $k(x_i, x_j|\theta)$ with hyperparameters θ and transforms the inputs x utilizing

$$k(x_i, x_j|\theta) \to k(g(x_i, w), g(x_j, w)|\theta, w) \tag{1}$$

where g(x,w) is a non-linear mapping provided by a DNN, parametrized by weights w. Here we utilize a radial basis function (RBF) kernel

$$k_{RBF}(x, x') = exp(-\frac{1}{2}||x - x'||/l^2) \tag{2}$$

All hyperparameters of the base kernel, θ, and of the DNN, w, are learned jointly by maximizing the log marginal likelihood $\mathcal{L}$ of the GP.

$$\mathcal{L} = log(p(y|w, \theta, X)) \propto -[y^T(K_{w,\theta} + \sigma^2 I)^{-1}y + log[K_{w,\theta} + \sigma^2 I]] \tag{3}$$

This is accomplished via the chain rule to compute derivatives of $\mathcal{L}$ with respect to w

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial \mathcal{L}}{\partial K_{w,\theta}} \frac{\partial K_{w,\theta}}{\partial \theta}, \frac{\partial \mathcal{L}}{\partial w} = \frac{\partial \mathcal{L}}{\partial K_{w,\theta}} \frac{\partial K_{w,\theta}}{\partial g(x, w)} \frac{\partial g(x, w)}{\partial w} \tag{4}$$

The derivative of $\mathcal{L}$ with respect to the n x n covariance matrix $K_{w,\theta}$ is given by

$$\frac{\partial \mathcal{L}}{\partial K_{w,\theta}} = \frac{1}{2}(K_{w,\theta}^{-1} y y^T K_{w,\theta}^{-1} - K_{w,\theta}^{-1}) \tag{5}$$

To avoid the poor scaling of the $K_{w,\theta}^{-1}$ calculation, the KISS-GP scaled kernel interpolation procedure developed by Wilson and Nickisch is used.[60] Utilization of the KISS-GP approximation enables linear scaling in the number of training data. Following hyperparameter optimization on a training data set, X, the predictive distribution of the GP evaluated at the test points X_* is given by the standard GPR expressions:

$$f_*|X_*, X, w, \theta, \sigma^2 \sim N(E[f_*], cov(f_*)) \tag{6}$$

$$E[f_*] = \mu_{X_*} + K_{X_*,X}[K_{X,X} + \sigma^2 I]^{-1} y \tag{7}$$

$$cov(f_*) = K_{X_*,X_*} - K_{X_*,X}[K_{X,X} + \sigma^2 I]^{-1} K_{X,X*} \tag{8}$$

All DKL calculations are implemented through the PyTorch based GPU accelerated GPyTorch library.[61] The calculations are performed on a single NVIDIA A100 GPU. For DKL we use a four hidden layer neural network featurizer with exponential linear unit activations and batch normalization at each layer feeding into a RBF kernel. The Adam optimizer[62] is applied to tune DKL weights. The width of neural network layers, number of training iterations, and learning rate are further tuned using Bayesian optimization for a randomly drawn data set from the 1000K flexible data set. The list of hyperparameters, optimized architecture, and search range are provided in Table S2.

Active Learning

Whereas previous ECG models have proven effective, they are limited by the need for a large number of QC calculations for parameterization. To overcome this, we implement three different AL query strategies to develop reduced size training data sets exhibiting negligble loss in performance relative to larger data sets.[39] The AL procedure is as follows: (i) a small random subset of data are drawn from the large unlabeled pool of CG configurations to initialize the training data set, (ii) the DKL-ECG model is trained on the current training data set, (iii) the trained DKL-ECG model is applied to all data remaining in the unlabeled pool, (iv) a query function is used to calculate a score for each sample in the pool, (v) the sample(s) with the best score are selected for labeling, labeled, and then added to the training data set, and then the cycle repeats from (ii) until convergence. It is important to note that there are a broad range query strategies available in the literature. Here, we have judiciously selected three: RQ, UQ, and EMOC query. In what follows, we approximate the functional map $\mathfrak{F} : \mathfrak{D} \rightarrow \mathfrak{Y}$ where $x \in \mathfrak{D}$ are the input features and $y \in \mathfrak{Y}$ are continuous output labels using DKL-ECG.

Random Query (Q_{ran}):

Random query serves as a baseline method for comparison against more sophisticated AL query strategies. Random query simply involves selecting a random subset of data from the unlabeled pool without any scoring system for the unlabeled data.

Uncertainty Query (Q_{unc}):

Uncertainty query is one of the most commonly used AL strategies. Here, the mean and uncertainty derived from the DKL regression model are used to construct a score that balances exploration and exploitation defined by:[63]

$$Q_{unc}(\mathfrak{U}) = \underset{x^{(i)} \in \mathfrak{U}}{\operatorname{argmin}} \frac{|\mu(x^{(i)})|}{\sqrt{\sigma^2(x^{(i)})}}, \tag{9}$$

The numerator of the query function, the absolute predictive mean, is exploitative whereas the denominator, the predicted variance, is explorative.

EMOC query (Q_{emoc}): As an advanced balance between exploitation and exploration, we employ the EMOC AL query algorithm. EMOC avoids irrelevant or redundant samples being considered for labeling that would not lead to improvement in model output after retraining. The EMOC based AL strategy for GP classification tasks was proposed by Freytag et al.,[50] with Käding et al.[51] extending the approach to regression problems. The approach has not been widely adopted in the chemical community due to its cumbersome mathematical form requiring transformation to be performed on the kernel elements. Here we reproduce the main elements of the derivation of the EMOC strategy in closed form and provide a generic numerical implementation that can be applied to GPR methods based on a highly customizable and GPU accelerated PyTorch framework.[64] This implementation is available in our GitHub repository [1].

EMOC selects the samples from the unlabeled pool $\mathcal{U}$ with the largest expected change of model output. For a new model $\mathfrak{F}'$ obtained by updating an old model $\mathfrak{F}$ with $(x^{'}, y^{'})$,

$$Q_{emoc}(\mathfrak{U}) = \underset{x^{(i)} \in \mathfrak{U}}{\operatorname{argmax}} \Delta \mathfrak{F} = \underset{x^{(i)} \in \mathfrak{U}}{\operatorname{argmax}} \mathtt{E}_x \mathtt{E}_{y^{'}|x^{'}} ||\mathfrak{F}'(x) - \mathfrak{F}(x)||_P. \tag{10}$$

To derive the EMOC criterion described in eqn. 10 for GPR-like methods we apply a zero mean assumption to eqn. 7 and assume an arbitrary kernel function $\mathcal{K}(.,.)$, from which the prediction by $\mathfrak{F}$ can be written as

$$\mathfrak{F}(x) = \sum_j \alpha_i \mathcal{K}(x^{(j)}, x) = k(x)^T \alpha, \tag{11}$$

where α is the weight vector resulting from the GPR training and $x^{(j)} \in L$, the pool of labelled samples used for training the model. Now we compute the EMOC criterion for GPR as,

$$\Delta \mathfrak{F}(x^{'}) = \mathtt{E}_x \mathtt{E}_{y^{'}|x^{'}} ||k^{'}(x)^T \Delta \alpha||_P, \tag{12}$$

where $k^{'}(x) = [k(x), \mathcal{K}(x^{'}, x)]$ and $\Delta\alpha$ indicate the difference of the current model weight α from the new model weights obtained by the addition of the new labelled samples, $(x^{'}, y^{'})$. There exists a closed form for $\Delta\alpha$ for GPR which is given by,[50]

$$\Delta \alpha = \frac{k(x^{'})^T \alpha - y^{'}}{\sigma_n^2 + \sigma_{f'}^2(x^{'})} \begin{bmatrix} (\mathcal{K} + \sigma_n^2 I)^{-1} k(x^{'}) \\ -1 \end{bmatrix}, \tag{13}$$

[1] https://github.com/TheJacksonLab/ECG_ActiveLearning

where $\sigma^2_{f'}$ corresponds to the predictive variance of $x^{'}$.

Eqn. 13 has a dependence on $y^{'}$. But $y^{'}$ is not known *a priori* before the labelling, hence a marginalization over $y^{'}$ must be performed. For this purpose we decompose eqn. 13 based on its dependence on $y^{'}$ as follows:

$$\Delta\alpha = g(x^{'})(k(x^{'})^T\alpha - y^{'}). \tag{14}$$

We will now rewrite eqn. 12 to derive EMOC for new samples $x^{'}$ with respect to a single training sample x:

$$\Delta\mathfrak{F}(x^{'}, x) = \mathtt{E}_{y^{'}|x^{'}}||k^{'}(x)^T g(x^{'})(k(x^{'})^T\alpha - y^{'})||_P. \tag{15}$$

Let $\nu = ||k^{'}(x)^T g(x^{'})||$, $c = k(x^{'})^T\alpha$ and substituting $z = y^{'} - c$ in the above equation results in

$$\Delta\mathfrak{F}(x^{'}, x) = ||\nu||_P \int_{\mathfrak{Y}} ||z||_P \mathtt{p}(z + c|x^{'})dz. \tag{16}$$

We exploit that $\mathfrak{F}$ is modeled by a GP. Hence our posterior distribution can be approximated as Gaussian as follows:

$$\mathtt{p}(z + c|x^{'}) = \mathcal{N}(z + c|\mu(x^{'}), \sigma^2_f(x^{'})).$$

This leads to

$$\Delta\mathfrak{F}(x^{'}, x) = ||\nu||_P \mathtt{E}[||z||_P], \tag{17}$$

There exists a closed form solution for the non-central Pth-moment of the Gaussian distribution given by ,

$$\mathtt{E}[||z||_P] = \sigma^P.2^{\frac{P}{2}}.\frac{\Gamma(\frac{1+P}{2})}{\sqrt{\pi}}.{}_1F_1\left(-\frac{P}{2},\frac{1}{2},-\frac{1}{2}\left(\frac{\mu(x')}{\sigma_f^2(x')}\right)^2\right), \tag{18}$$

where $\Gamma(.)$ is the gamma function. The confluent hypergeometric function ${}_1F_1(.,.,.)$ is given by

$${}_1F_1(a,b,z) = \sum_{n=0}^{\infty}\frac{a^{(n)}z^n}{b^{(n)}n!}. \tag{19}$$

The computational workflow is illustrated in Figure 1. In this work we use eight distinct data sets that represent rigid and flexible MD runs performed at four temperatures (50K, 300K, 600K, 1000K). The AL methods above are first applied independently to all distinct data sets. Each data set contains 9,700 configurations wherein the first 8,500 are used for each AL query, the next 200 are skipped to ensure decorrelation between training and validation sets, and the final 1,000 are used for independent validation. The labeled pool of configurations for every AL run is initialized with three random samples. The AL query process is run for 1,500 iterations, with the highest scored configuration from the unlabeled pool being added to the training data set; in the case of random sampling, a randomly selected data point from the unlabeled pool is added. The query width in all cases is 6,995. All AL learning curves (SI Figure S7) are averaged over five independent runs initialized with different random seeds. We measure the performance of the AL methods by reporting the root mean square error (RMSE) and coefficient of determination (R^2) associated with predicting the HOMO energy of each configuration. To further test the robustness of the AL query strategies applied across multiple temperatures we create temperature aggregated rigid and flexible data sets, resulting in a large $\mathfrak{U}$ consisting of an initial sample size of 34,000 (SI Figure S6). A query width of 32,495 is used. Four different test data sets each consisting of 1000 samples drawn at different temperature are kept for validation. Finally, to test the generalization of the DKL-ECG model outside of training data sets, we also generated a global validation data set using LHS.[65] LHS is used to sample 2,000 dihedral configurations each for rigid and flexible.

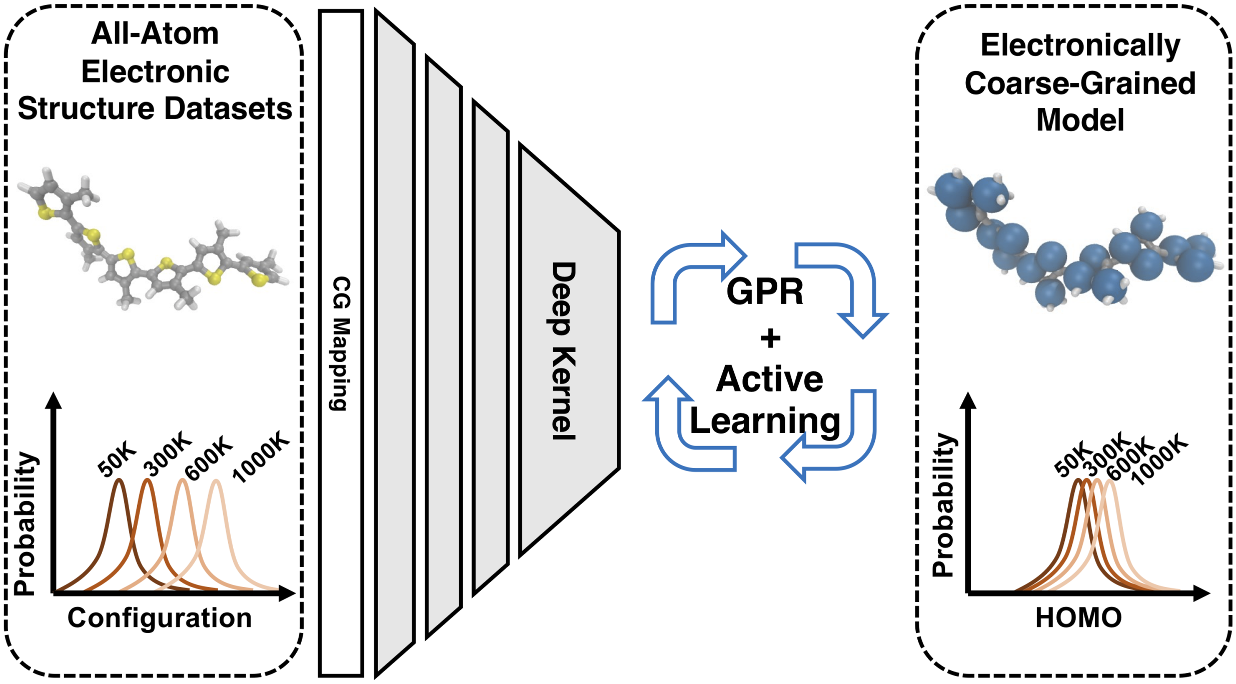

Figure 1: Computational approach for DKL-ECG used in this work. All-atom configurations are mapped to CG representations. CG distance matrices are transformed by DNN to generate a low dimensional latent representation. The intermediate latent representation drives kernel based regression and the AL of the HOMO energies.

Results and Discussion

Energetic and Configurational Distributions

We begin by analyzing the statistics of the input (CG distance matrices) and output (HOMO energies) distributions for each data set. The HOMO energy distributions for all flexible and rigid data sets are shown in the SI Figure S2 and Figure S3. The HOMO energy distributions of the rigid data sets exhibit lower values of the distribution mean (50K: -7.250 eV, 300K: -7.404 eV, 650K: -7.452 eV, 1000K: -7.451 eV) relative to the flexible data sets (50K: -7.150 eV, 300K: -7.330 eV, 650K: -7.354 eV, 1000K: -7.325 eV), which is attributable to the bias induced by the intrathiophene degrees of freedom being frozen at their minimum energy geometry in the rigid simulations. The mean HOMO energies for both rigid and flexible shift to higher energy as the temperature is lowered, though the maximum observable HOMO energy increases with increasing temperature, consistent with the ability of high temperature simulations to access planar configurations in which intermonomer couplings are largest. Also consistent with the difference between the rigid and flexible data sets is the larger value of the standard deviation (SI Table S1) of the HOMO energy in the flexible data sets (50K: 0.072 eV, 300K: 0.195 eV, 650K: 0.231 eV, 1000K: 0.253 eV) relative to the rigid data sets (50K: 0.078 eV, 300K: 0.170 eV, 650K: 0.202 eV, 1000K: 0.214 eV). The lower value of the standard deviation of the rigid data sets relative to the flexible data sets is attributable to the lack of fluctuations of the intrathiophene degrees of freedom in the rigid data set, which reduces the diversity of conformations, thus reducing the width of the HOMO energy distribution. For both rigid and flexible data sets, the structure of the distribution shifts dramatically in moving from 50K to 300K, indicating the activation of conformational degrees of freedom with strong electronic coupling that are frozen out in the 50K simulations.

We next characterize the data set dependence of the configurational space used as input to DKL-ECG via analysis of the intermonomer dihedral distributions (SI Figure S4 - S5). All five S3MT intermonomer dihedrals are histogrammed and averaged into a single effective dihedral degree of freedom, the free-energy surface of which is obtained via Boltzmann Inversion (SI Figure S4-S5). The average flexible dihedral barrier heights at 50K, 300K, 650K, and 1000K are $\sim 10kcal/mol$, $\sim 2kcal/mol$, $\sim 1kcal/mol$, and $\sim 0.7kcal/mol$, respectively. The average rigid dihedral barrier heights at 50K, 300K, 650K, and 1000K are $\sim 10kcal/mol$, $\sim 4kcal/mol$, $\sim 2kcal/mol$, and $\sim 2kcal/mol$, respectively. Rigid dihedral barriers are on average larger than the corresponding flexible dihedral barriers due to the rigidity constraints on the monomers which prevent the rotation or bending of the methyl group upon dihedral rotation.

To further characterize conformational diversity across all data sets we train a DKL-ECG model and extract the final layer of the DNN (Figure 1). DKL-ECG is trained on a random subset of ~ 1000 configurations drawn from both 1000K rigid and flexible data sets; the configurations drawn from the 1000K rigid and flexible data sets represent the conformationally most diverse data due to possessing the lowest free energy barriers in the system (SI Figure S4-S5). The final layer of the DNN contains a compressed feature representation of the high dimensional CG distance matrix input (Figure 1). We use the weights of the DKL-ECG model trained on the 1000K data subset and pass all 34K training samples through the network to extract the corresponding latent vectors. To aid in visualization, PCA is applied to the 21-dimensional latent space representation and the eigenvectors of the two largest principal components are extracted and plotted for all rigid (Figure 2A) and flexible (Figure 2B) data sets as a function of temperature. To qualitatively understand the conformational diversity spanning the principal components of the latent space, configurations are visualized using Ovito[66] (Figure 2).

Analysis of the principal components of the DKL-ECG latent space shows a broad diversity of conformations sampled at different temperatures. As anticipated from trends in the dihedral free-energy surfaces, increasingly diverse conformational states are sampled at higher temperature simulations for both rigid and flexible systems. Notably, all cis conformations of the intermonomer dihedrals are only obtained at the highest temperature simulations (1000K) with the sparsest sampling. 300K, 650K, and 1000K simulations exhibit broad overlap of sampled conformations, consistent with the fact that dihedral degrees of freedom are well sampled at even 300K for the timescale of all simulations. Notable in these plots is the strongly localized nature of conformational space sampled in 50K MD simulations, as represented by the islands of configurational space within the Figure 2 plots. From these results, it is anticipated that training exclusively on 50K configurations can induce low prediction transferability due to the lack of sampling other, more diverse configurations. We further plot the HOMO energy dependence of the DKL-ECG latent space in SI Figure S6, observing only weak structuring of the latent space according to HOMO energy prediction.

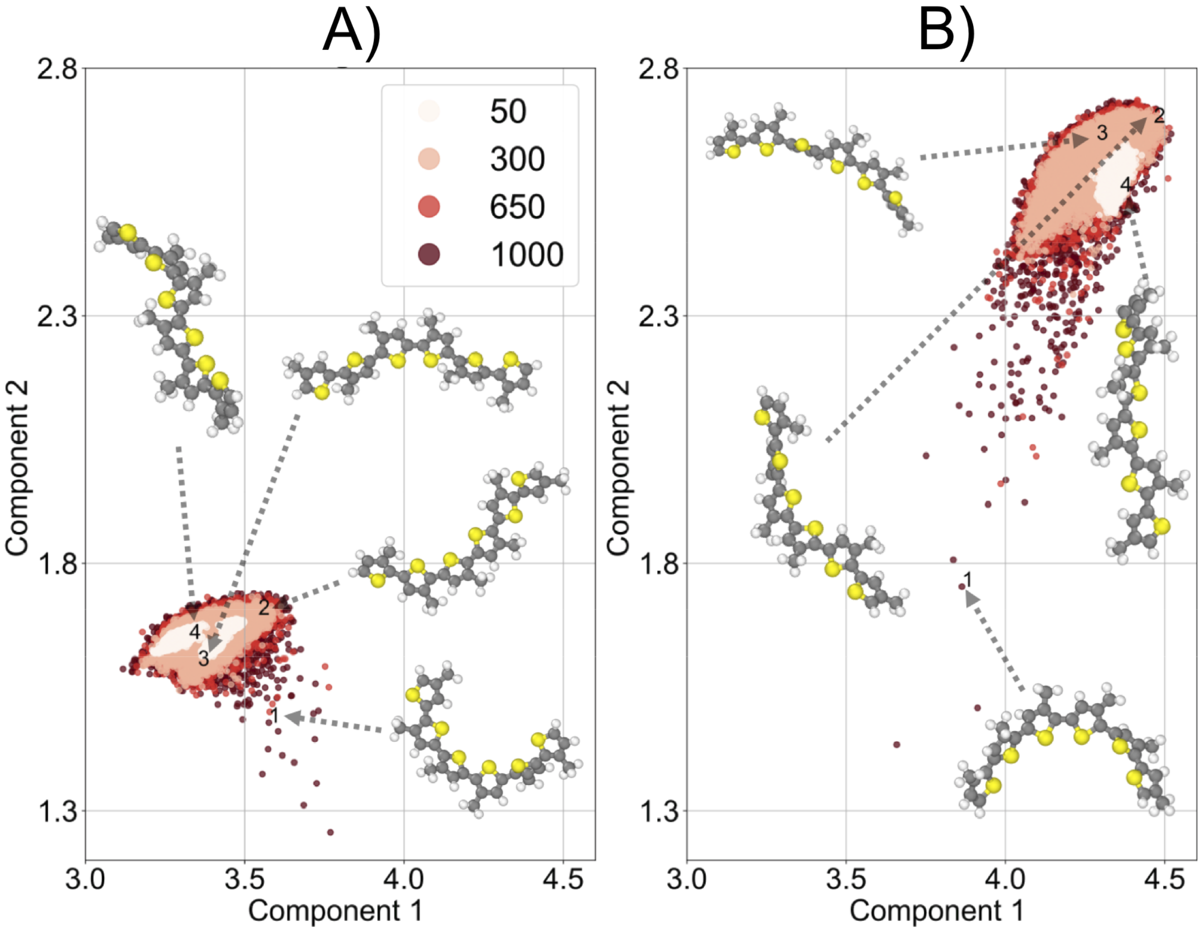

Figure 2: Visualization of the two largest principal components of the latent space representations of the CG distance matrices for (a) rigid and (b) flexible simulations across all sampled temperatures. The legend refers to reference temperatures (K) of the samples. Representative configurations are shown as insets.

Multiple Temperature DKL-ECG Performance

With an understanding of the conformational spaces sampled using different temperature MD simulations, we next examine the accuracy and transferability of DKL-ECG across all temperatures. These DKL-ECG models represent the highest-quality electronic structure data used in ECG predictions to date (range-separated, dispersion-corrected hybrid DFT). In an attempt to minimize the quantity of training data required to train DKL-ECG at each temperature, we first examine the single-temperature performance when using three different AL queries: RQ, UQ, and EMOC. In all cases AL algorithms are run for 1,500

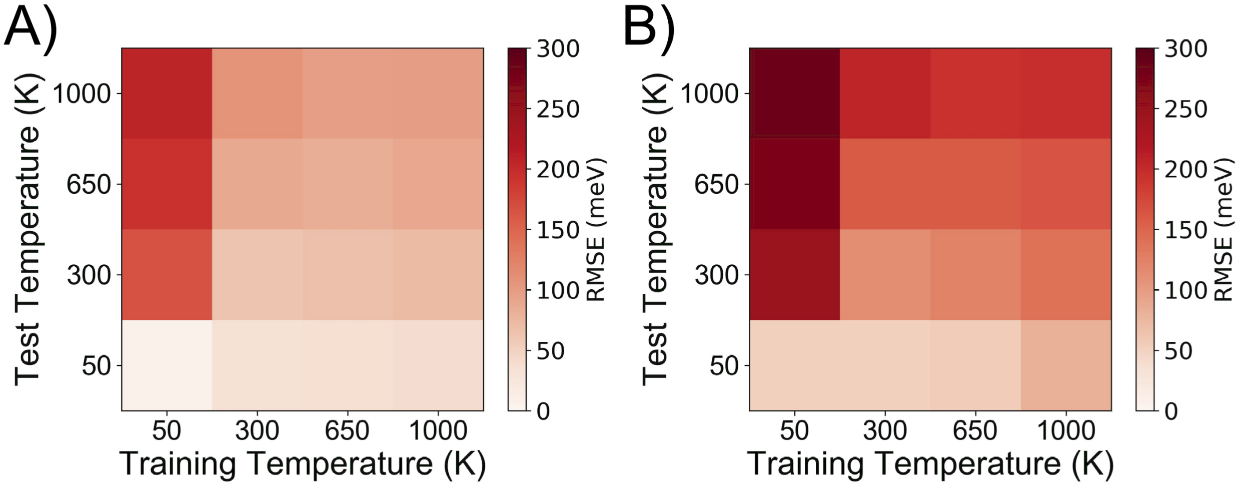

Figure 3: Heatmap demonstrating the transferability of single temperature DKL-ECG models trained using EMOC for A) rigid B) flexible data sets. The colorbar indicates model RMSE error in meV for held out test data sets. Axes labels indicate the reference temperature of the training and test data sets.

cycles ($\approx$ 18% of 8500 single temperature samples) until training data sets of size 1,500 are assembled. Reported performance of a single temperature model occurs by application to a held-out 1000 sample test test for each temperature model (e.g. a model trained with a data set drawn for flexible at 1000K is trained for 1500 AL cycles and validated on held out test sets of 1000 samples each at 1000K, 650K, 300K and 50K). As noted in the methods, all the metrics are averages over five independent runs of the AL queries. All quantitative results of these computational experiments are listed in SI Tables S3-S10.

Surprisingly, RQ, UQ, and EMOC all exhibit similar quantitative performance for single temperature prediction tasks. For rigid simulations in which training and testing occur at the same temperature (SI Tables S3-S6), R^2 of $\sim 0.8 - 0.9$ is achieved using only 1,500 training points, with RMSE ranging from 10-100 meV depending on temperature. Test errors (RMSE) correlate positively with temperature, consistent with a broader feature space that must be learned. Similar qualitative results are obtained for flexible DKL-ECG (SI Tables S7-S10), with R^2 of $\sim 0.3 - 0.6$ and RMSE ~50-300 meV. These results are competitive with those of previous work utilizing larger data sets ($\sim 10^4$) and a lower level of theory. While the equivalent performance of all AL query strategies might seem strange, this is understood by considering the nature of the training data. In the case of single temperature training, all conformations are drawn from the same Boltzmann distribution function and are statistically uncorrelated. In this scenario, randomly sampling this distribution function appears to be approximately as effective as more sophisticated AL algorithms such as UQ and EMOC. However, in cases in which the AL algorithm is sampling data drawn from distinct distributions, one will expect more dramatic effects, a task we explore later in this work for multiple temperature DKL-ECG model training.

To visualize the temperature transferability of DKL-ECG models trained on single temperatures, the results of the computational experiments (SI Tables S3-S10) using the EMOC query algorithm are summarized as a heatmap in Figure 3. For both rigid and flexible data sets, DKL-ECG models trained at higher temperatures show strong transferability to lower temperature data sets. For example the rigid model trained at 1000K shows RMSE of 98, 89, 71, and 39 meV validated at 1000, 650, 300 and 50K respectively. While the rigid model trained at 50K performs well at 50K, its performance decays strongly with increasing temperature of the validation data set, exhibiting the highest RMSE $>$ 200 meV at 1000K. A similar trend of model transferability can be observed for flexible data sets in Figure 3B) in which DKL-ECG models trained at high temperatures are generally transferable to lower temperature data sets, but not the reverse. We attribute the transferability of high temperature DKL-ECG models to lower temperature data sets to the broad configurational space sampled by higher temperature MD simulations as shown in Figure 2. Interestingly, the performance of high temperature models applied to low temperature data sets is often quantitatively comparable to the performance of the DKL-ECG model trained at the targeted prediction temperature (See SI Tables S3-S10). This leads to the conclusion that training DKL-ECG models on higher temperature configurations is generally a robust strategy even for predicting electronic structure on configurations sampled from significantly lower temperature MD simulations. While this appears to be a reliable rule-of-thumb, pathological cases can be observed for flexible data sets at very high temperature differences between training and test (e.g. Table S7), as evidenced by the low R^2 values.

We next examine the ability of AL queries to construct training data sets for the purpose of multiple temperature DKL-ECG predictions. As noted in the methods section, combining the four single temperature data sets creates a training space of 34,000 samples. Similar to the previous section, AL algorithms are run for 1,500 cycles, but on a much larger and diverse sampling space (≈4.4% of 34000 multi-temperature samples) until training data sets of size 1,500 are assembled. The AL mixed temperature models are validated for each of the 1000 test samples drawn at each temperature to probe model performance and transferability.

We begin by analyzing the regions of the aggregated (34k) feature space from which each of the AL queries samples when assembling the 1,500 training data points (Figure 4). Only EMOC effeectively samples from diverse regions of the configurational distribution provided the low number of training data points relative to the full data set. This advantage becomes especially apparent in the isolated tail points corresponding to high temperature regions (Figure 4C and 4G). UQ performs worse than EMOC in this task, but slightly better than RQ, which samples a very localized region of configuration space. We further investigate the configurational diversity sampled by EMOC by examining the effect of different random initialization on the AL query. This is accomplished by linking picked samples to their corresponding temperatures from which they were drawn. The number of unique samples picked at each temperature for each AL query is shown in SI Table S11. It is observed that, averaged over five independent runs of the AL queries, EMOC samples significantly fewer unique configurations than both RQ and UQ. This is consistent with the design of EMOC in which samples are picked to maximize the expected model output change (eqn. 10), and consequently should be more robust to random initialization of the AL query.

EMOC is also observed to preferentially sample configurations from higher temperature data sets which exhibit more conformational diversity (Figure 5). It is clear from Figure 5 that EMOC can identify the lower configurational diversity of the 50K MD simulations, and thus reduces the number of configurations drawn from these data sets. Similarly, EMOC pulls the most configurations from the 1000K MD simulations, as the feature space (Figure 2) and output (SI Figure S2 and S3) of the 1000K configurations exhibit the most diversity. While the different data sets all exhibit some finite overlap with each other due to sampling being governed by the Boltzmann distribution, EMOC can intelligently navigate the differences between distribution functions to ensure that a broad feature space is being sampled, which we anticipate to be an even more desirable feature of EMOC for chemical systems in which sampling is not all drawn from overlapping distribution functions (e.g. chemical space).

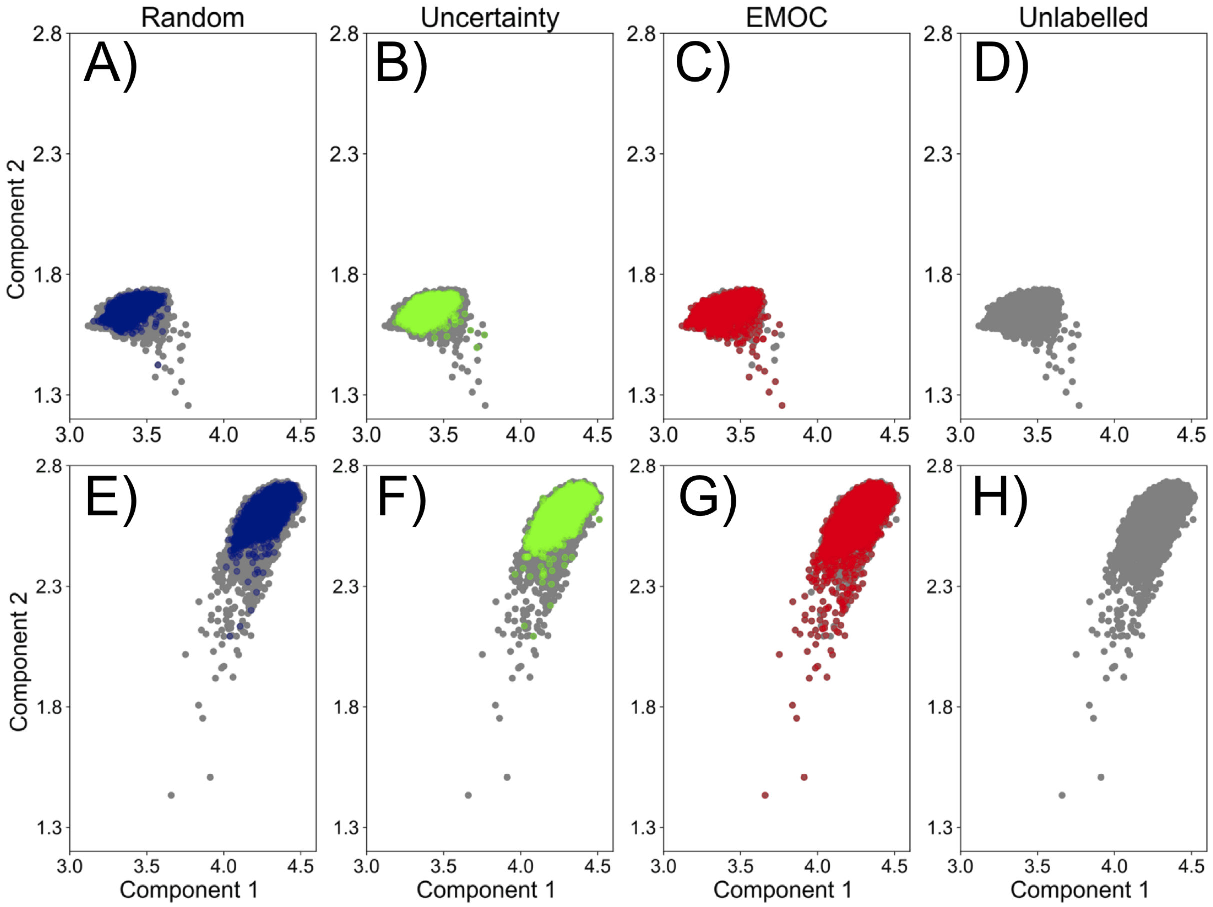

Figure 4: PCA Visualization of the DKL latent space representation of the entire CG configurational sampling space accessible via AL, grouped by AL query strategy for (a-d) Rigid (e-f) Flexible. Titles refers to AL query strategy and the right most panels show the entire unlabelled sampling space for comparison. The color of the points corresponds to the AL query method used for sampling. Results are aggregated over all five independent AL runs.

With an understanding of the different feature space sampling of the AL queries, we

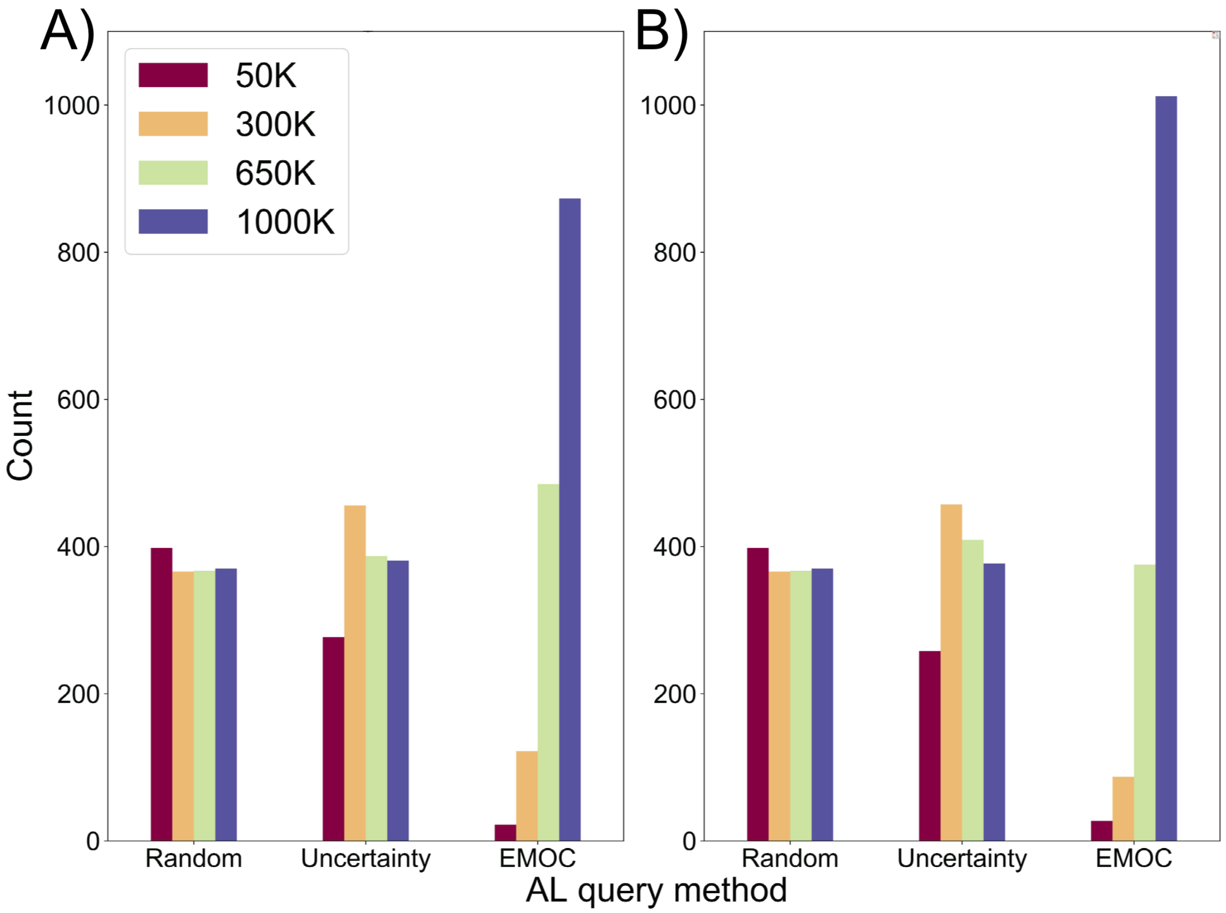

Figure 5: Histogram showing the fraction of samples drawn by each of the AL query strategies at each temperatures used for constructing the mixed temperature DKL-ECG model. (a) Rigid (b) Flexible.

examine the quantitative accuracy of DKL-ECG when trained on multiple temperature data sets. All reported results are drawn from held out 1,000 data point test sets taken from each data set. For DKL-ECG trained on rigid data sets across all temperatures (Figure 5), results are reported in Table 1. As for the single temperature results, all AL query methods lead to accurate DKL-ECG models. As a benchmark, previous studies employing a much simpler semi-empirical level of theory (ZINDO/S) obtained RMSE of $\sim 15meV$ and R^2 of 0.99 when training over 8,000 data points.[23,24] Here, the $\omega B97X - D3$ data set presents a significantly more challenging data set from which to learn electronic structure correlations due to the presence of non-local electron correlation and exact exchange not present in semi-empirical methods. The best performing DKL-ECG model at 300K resulting from UQ exhibits RMSE of $\sim 65meV$ and R^2 of 0.84 using only 1500 training datapoints, demonstrating the power of the DKL-ECG approach. Exceptional performance ($\sim 10meV, R^2 = 0.96$) is obtained over the reduced conformational space of the 50K data set. Performance decays on higher temperature data due to the broader configurational space of these data sets.

While the predictive accuracy of the DKL-ECG methods for all AL query methods is satisfactory, the most surprising result is that RQ, UQ, and EMOC AL strategies all perform nearly identically despite the greater sophistication of UQ and EMOC. To understand this result, we reexamine the PCA configuration space distributions shown previously in Figure 2. It is well known that AL algorithms such as UQ and EMOC are most effective when sampling data distributions existing in distinctly different basins.[51] The PCA distribution plots of Figure 2A) show that the rigid data set configurations exist in effectively only two distinct basins - one predominantly populated by 300K, 650K, and 1000K and the other predominantly populated by 50K. Consequently, we hypothesize that the reason for the equivalent performance of all three AL algorithms is due to the fact that each method is sampling from smooth, well-connected basins of conformational space when sampled from individual temperatures distributions. In this case, randomly sampling these distributions using RQ is approximately as effective as UQ or EMOC approaches. However, we anticipate that in systems sampling distinctly different conformational distributions, the more advanced UQ and EMOC algorithms should outperform RQ. Having said that, we should mention that our PyTorch based implementation of EMOC should be broadly useful to the community beyond the scope of problems discussed here.

Next RQ, UQ, and EMOC are applied to all flexible data sets with the results shown in Table 1. Similar to the rigid case, all models produce satisfactory RMSE and R^2 values across all temperatures. As a benchmark against previous semi-empirical results at 300K (RMSE $\sim 90meV$, $R^2 \sim 0.57$) trained on 8,000 datapoints, our 1,500 datapoint best AL results with UQ exhibit comparable RMSE $\sim 122.4meV$, and a slightly improved $R^2 \sim 0.62$.[23,24] The flexible data set also has more conformational diversity compared to the rigid, worsening the performance due to configurational degeneracy induced by the CG mapping operator. To further explore any potential advantages of the more sophisticated EMOC query strategy compared to RQ and UQ we examine the evolution of the RMSE over the entire AL training cycle by plotting learning curves in SI Figure S7. For rigid and flexible data sets, all query strategies exhibit similar performance, with RQ and UQ significantly outperforming EMOC at low temperatures. However, in flexible data sets we note that EMOC significantly outperforms RQ and UQ at 1000K, while degrading strongly as the temperature is lowered. While this result is primarily exploratory, it is consistent with the hypothesis that EMOC will outperform RQ and UQ in situations when data is drawn from distinctly different basins. The flexible, 1000K data set represents the most conformationally diverse data set of all those sampled in this work via MD simulations. Consequently, the consistently higher performance of EMOC relative to RQ and UQ for the high temperature flexible data sets suggests the hypothesis that it should outperform RQ and UQ if samples are drawn from even more diverse configuration sets.

One final note regarding the performance similarities of different AL queries concerns the overlap of the HOMO energy output distributions shown in SI Figure S2 and S3. In all cases, regardless of features, the predicted output distributions exhibit significant overlap. We suspect that this is yet another confounding reason for the similar quantitative performance of all AL methods here; not only do the basins from which the feature space is sampled strongly overlap, but the output distributions are also quite similar. Moreover, as the dihedrals primarily govern the HOMO energy value, dramatically different configurations (e.g. all cis vs all trans) could exhibit similar HOMO energies if all dihedrals are planar in different cases. While the primary purpose of this work is the parameterization of DFT-quality DKL-ECG models across multiple temperatures, we anticipate that further improvement and refinement via the use of AL queries will be obtained in systems in which large changes in configurational space typically correspond to large changes in the output electronic property of interest.

As a final validation of AL parameterized DKL-ECG, we examine the performance of the mixed temperature DKL-ECG model applied to a data set in which configurations are drawn from harmonically constrained diherals in conjunction with latin hypercube sampling. As the dihedral distribution of this data set is non-canonically sampled, it represents a useful test case for the transferability of the multiple temperature DKL-ECG model. RMSE and R^2 are shown in Table 1, with confidence intervals and their statistics reported in SI Figure S8 and SI Table S12. It is seen that all DKL-ECG prediction means fall within the 95% confidence intervals, showing that there is minimal overfitting or outliers present in the DKL-ECG prediction model. Performance on the LHS data obtained RMSE of 57.6 meV and R^2 of 0.88 and RMSE of 99.3 meV and R^2 of 0.68 for rigid and flexible data sets, respectively. The average confidence intervals associated with these predictions are $\sim 2-5meV$, showing excellent quantitative performance of the DKL-ECG model across a conformationally diverse, unseen validation data set. The DFT evaluation of single configurations requires 240 s whereas the DKL-GPR performed prediction for 2,000 LHS data set in $\sim$ 0.006s thereby showcasing an impressive $\sim 10^7$ relative speed up.

Finally, to provide a sense of the absolute AL DKL-ECG accuracy, we compare the performance to a DKL-ECG model trained on the full 34,000 configuration data set (Table 2). The DKL-ECG model that utilizes 4.4% of the total data achieves impressively comparable R^2 values for the rigid data set, though similarly exceptional results are not apparent for the flexible data sets. Consequently, it appears that 1,500 judiciously selected data points is sufficient to sample all relevant dihedral degrees of freedom in the rigid data set, though the additional degrees of freedom present in the flexible data set present challenges as the CG mapping operator degeneracy associated with these degrees of freedom induces an intrinsic noise on the prediction task. However, Table 2 drives home the critical point that, with enough data, electronic prediction models acting on entirely CG molecular representations are capable of obtaining DFT-quality electronic structure predictions without recourse to all-atom backmapping and *ad nausuem* QC calculations.

Table 1: EMOC DKL. Metrics averaged over five independent runs.

Temperature (K)	Rigid		Flexible	
	RMSE (meV)	R^2	RMSE (meV)	R^2
1000	99.3 (±) 2.8	0.78	202.5 (±) 3.1	0.34
650	88.1 (±) 2.2	0.81	168.5 (±) 3.7	0.43
300	69.5 (±) 2.0	0.82	136.5 (±) 2.0	0.52
50	15.7 (±) 2.3	0.91	58.1 (±) 17.8	0.49
LHS (300)	57.6 (±) 1.1	0.88	99.3 (±) 3.7	0.68

Table 2: DKL model trained on the entire sampling scape of 34,000 configurations. Metrics averaged over five independent runs.

Temperature (K)	Rigid		Flexible	
	RMSE (meV)	R^2	RMSE (meV)	R^2
1000	43.4 (± 0.7)	0.96	164.9 (± 3.3)	0.56
650	34.7(± 0.6)	0.97	129.4 (± 1.4)	0.67
300	22.5 (± 0.3)	0.98	85.3 (± 0.83)	0.81
50	3.9 (± 0.4)	0.99	36.0 (± 4.6)	0.74
LHS (300)	21.7 (± 0.3)	0.98	55.8 (± 2.6)	0.90

Conclusions

We have constructed an ECG prediction model within the framework of DKL. Using this framework, we have developed DKL-ECG models at the highest level of electronic structure theory to-date, range separated hybrid DFT, which increases the applicability of ECG to modeling real molecular systems with high accuracy. The ability of DKL-ECG to provide uncertainty estimates on predictions has facilitated the incorporation of AL algorithms for efficiently training DKL-ECG models with minimal data and maximum configurational and temperature transferability. While the EMOC AL algorithm clearly samples a more diverse feature space than the RQ or UQ algorithms, suggesting improved prediction accuracy and transferability, predictive performance among all three AL query methods is effectively identical. We attribute this result to the redundancy of Boltzmann sampling for the molecule of interest, and anticipate that EMOC performance should be significantly improved over RQ and UQ in future applications that sample data from strongly different distributions. The results presented in this work significantly advance the ECG method, providing a means of performing DFT-accurate electronic predictions directly from CG representations.

Acknowledgement

This material is based upon work supported by Laboratory Directed Research and Development (LDRD-CLS-1-630) funding from Argonne National Laboratory, provided by the Director, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-06CH11357. N.E.J acknowledges support from the Dreyfus Program for Machine Learning in the Chemical Sciences and Engineering during this project. This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357. Argonne National Laboratory's work was supported by the U.S. Department of Energy, Office of Science, under contract DE-AC02-06CH11357. We gratefully acknowledge the computing resources provided on Bebop and Swing, high-performance computing clusters operated by the Laboratory Computing Resource Center at Argonne National Laboratory. G. S. would like to thank Dr. Murali Emani and Greg Pauloski for useful discussions on PyTorch compilation.

Supporting Information Available

The Supporting Information is available free of charge on the ACS Publications website. Supporting information contains: Comparison of DKL-ECG and Traditional GPR, HOMO Energy Statistics, Intermonomer Dihedral Statistics, Coarse-Grained Mapping Operator, Representative S3MT All-Atom Configuration, Hyperparameters for DKL-ECG, Visualization of the DKL-ECG Feature Space, Single Temperature DKL-ECG Performance, Learning Curves for Multiple Temperature DKL-ECG, Unique Samples from Multiple Temperature DKL-ECG AL Queries, and Multiple Temperature DKL-ECG Confidence Intervals on LHS Validation.

References

(1) Wasielewski, M. R. Self-Assembly Strategies for Integrating Light Harvesting and Charge Separation in Artificial Photosynthetic Systems. *Accounts of Chemical Research* **2009**, *42*, 1910–1921, Publisher: American Chemical Society.

(2) Petersen, M. K.; Voth, G. A. Characterization of the Solvation and Transport of the Hydrated Proton in the Perfluorosulfonic Acid Membrane Nafion. *The Journal of Physical Chemistry B* **2006**, *110*, 18594–18600, Publisher: American Chemical Society.

(3) Burgess, M.; Moore, J. S.; Rodríguez-López, J. Redox Active Polymers as Soluble Nanomaterials for Energy Storage. *Accounts of Chemical Research* **2016**, *49*, 2649–2657, Publisher: American Chemical Society.

(4) Nelson, J.; Kwiatkowski, J. J.; Kirkpatrick, J.; Frost, J. M. Modeling Charge Transport in Organic Photovoltaic Materials. *Accounts of Chemical Research* **2009**, *42*, 1768–1778, PMID: 19848409.

(5) Alessandri, R.; Uusitalo, J. J.; de Vries, A. H.; Havenith, R. W. A.; Marrink, S. J. Bulk Heterojunction Morphologies with Atomistic Resolution from Coarse-Grain Solvent Evaporation Simulations. *Journal of the American Chemical Society* **2017**, *139*, 3697–3705, PMID: 28209056.

(6) Rühle, V.; Junghans, C.; Lukyanov, A.; Kremer, K.; Andrienko, D. Versatile Object-Oriented Toolkit for Coarse-Graining Applications. *Journal of Chemical Theory and Computation* **2009**, *5*, 3211–3223, PMID: 26602505.

(7) Rolland, N.; Modarresi, M.; Franco-Gonzalez, J. F.; Zozoulenko, I. Large scale mobility calculations in PEDOT (Poly(3,4-ethylenedioxythiophene)): Backmapping the coarse-grained MARTINI morphology. *Computational Materials Science* **2020**, *179*, 109678.

(8) Jackson, N. E. Coarse-Graining Organic Semiconductors: The Path to Multiscale Design. *The Journal of Physical Chemistry B* **2021**, *125*, 485–496, PMID: 33369413.

(9) Müller-Plathe, F. Coarse-Graining in Polymer Simulation: From the Atomistic to the Mesoscopic Scale and Back. *ChemPhysChem* **2002**, *3*, 754–769.

(10) Noid, W. G. Perspective: Coarse-grained models for biomolecular systems. *The Journal of Chemical Physics* **2013**, *139*, 090901.

(11) Reith, D.; Pütz, M.; Müller-Plathe, F. Deriving effective mesoscale potentials from atomistic simulations. *Journal of Computational Chemistry* **2003**, *24*, 1624–1636.

(12) Chaimovich, A.; Shell, M. S. Coarse-graining errors and numerical optimization using a relative entropy framework. *The Journal of Chemical Physics* **2011**, *134*, 094112.

(13) Izvekov, S.; Voth, G. A. A Multiscale Coarse-Graining Method for Biomolecular Systems. *The Journal of Physical Chemistry B* **2005**, *109*, 2469–2473, PMID: 16851243.

(14) Chakraborty, M.; Xu, J.; White, A. D. Is preservation of symmetry necessary for coarse-graining? *Physical Chemistry Chemical Physics* **2020**, *22*, 14998–15005.

(15) Noé, F.; Olsson, S.; Köhler, J.; Wu, H. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science* **2019**, *365*.

(16) Shmilovich, K.; Mansbach, R. A.; Sidky, H.; Dunne, O. E.; Panda, S. S.; Tovar, J. D.; Ferguson, A. L. Discovery of Self-Assembling -Conjugated Peptides by Active Learning-Directed Coarse-Grained Molecular Simulation. *The Journal of Physical Chemistry B* **2020**, *124*, 3873–3891, PMID: 32180410.

(17) Sidky, H.; Chen, W.; Ferguson, A. L. Molecular latent space simulators. *Chem. Sci.* **2020**, *11*, 9459–9467.

(18) Wang, W.; Gomez-Bombarelli, R. Coarse-graining auto-encoders for molecular dynamics. *npj Comput. Mater.* **2019**, *5*.

(19) Li, Z.; Wellawatte, G. P.; Chakraborty, M.; Gandhi, H. A.; Xu, C.; White, A. D. Graph neural network based coarse-grained mapping prediction. *Chemical science* **2020**, *11*, 9524–9531.

(20) Li, W.; Burkhart, C.; Polińska, P.; Harmandaris, V.; Doxastakis, M. Backmapping coarse-grained macromolecules: An efficient and versatile machine learning approach. *The Journal of Chemical Physics* **2020**, *153*, 041101.

(21) Harmandaris, V. A.; Adhikari, N. P.; van der Vegt, N. F. A.; Kremer, K. Hierarchical Modeling of Polystyrene: From Atomistic to Coarse-Grained Simulations. *Macromolecules* **2006**, *39*, 6708–6719.

(22) An, Y.; Deshmukh, S. A. Machine learning approach for accurate backmapping of coarse-grained models to all-atom models. *Chem. Commun.* **2020**, *56*, 9312–9315.

(23) Jackson, N. E.; Bowen, A. S.; Antony, L. W.; Webb, M. A.; Vishwanath, V.; de Pablo, J. J. Electronic structure at coarse-grained resolutions from supervised machine learning. *Sci. Adv.* **2019**, *5*.

(24) Jackson, N. E.; Bowen, A. S.; de Pablo, J. J. Efficient Multiscale Optoelectronic Prediction for Conjugated Polymers. *Macromolecules* **2020**, *53*, 482–490.

(25) Körzdörfer, T.; Brédas, J.-L. Organic Electronic Materials: Recent Advances in the DFT Description of the Ground and Excited States Using Tuned Range-Separated Hybrid Functionals. *Accounts of Chemical Research* **2014**, *47*, 3284–3291, PMID: 24784485.

(26) Rasmussen, C. E. Gaussian processes in machine learning. Summer school on machine learning. 2003; pp 63–71.

(27) Tovey, S.; Narayanan Krishnamoorthy, A.; Sivaraman, G.; Guo, J.; Benmore, C.;

Heuer, A.; Holm, C. DFT Accurate Interatomic Potential for Molten NaCl from Machine Learning. *The Journal of Physical Chemistry C* **2020**, *124*, 25760–25768.

(28) Li, Q.-J.; Küçükbenli, E.; Lam, S.; Khaykovich, B.; Kaxiras, E.; Li, J. Development of robust neural-network interatomic potential for molten salt. *Cell Reports Physical Science* **2021**, *2*, 100359.

(29) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical review letters* **2010**, *104*, 136403.

(30) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Physical Review B* **2013**, *87*, 184115.

(31) Deringer, V. L.; Bartók, A. P.; Bernstein, N.; Wilkins, D. M.; Ceriotti, M.; Csányi, G. Gaussian Process Regression for Materials and Molecules. *Chemical Reviews* **0**, *0*, null, PMID: 34398616.

(32) Denzel, A.; Kastner, J. Gaussian process regression for transition state search. *Journal of chemical theory and computation* **2018**, *14*, 5777–5786.

(33) John, S.; Csányi, G. Many-body coarse-grained interactions using Gaussian approximation potentials. *The Journal of Physical Chemistry B* **2017**, *121*, 10934–10949.

(34) Gkeka, P.; Stoltz, G.; Barati Farimani, A.; Belkacemi, Z.; Ceriotti, M.; Chodera, J. D.; Dinner, A. R.; Ferguson, A. L.; Maillet, J.-B.; Minoux, H., et al. Machine learning force fields and coarse-grained variables in molecular dynamics: application to materials and biological systems. *Journal of Chemical Theory and Computation* **2020**, *16*, 4757–4775.

(35) Mannodi-Kanakkithodi, A.; Pilania, G.; Huan, T. D.; Lookman, T.; Ramprasad, R. Machine learning strategy for accelerated design of polymer dielectrics. *Scientific reports* **2016**, *6*, 1–10.

(36) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *nature* **2015**, *521*, 436–444.

(37) Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J. V.; Lakshminarayanan, B.; Snoek, J. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *arXiv preprint arXiv:1906.02530* **2019**,

(38) Wilson, A. G.; Hu, Z.; Salakhutdinov, R.; Xing, E. P. Deep kernel learning. Artificial intelligence and statistics. 2016; pp 370–378.

(39) Sivaraman, G.; Krishnamoorthy, A. N.; Baur, M.; Holm, C.; Stan, M.; Csányi, G.; Benmore, C.; Vázquez-Mayagoitia, Á. Machine-learned interatomic potentials by active learning: amorphous and liquid hafnium dioxide. *npj Computational Materials* **2020**, *6*, 1–8.

(40) Doan, H. A.; Agarwal, G.; Qian, H.; Counihan, M. J.; Rodríguez-López, J.; Moore, J. S.; Assary, R. S. Quantum Chemistry-Informed Active Learning to Accelerate the Design and Discovery of Sustainable Energy Storage Materials. *Chemistry of Materials* **2020**, *32*, 6338–6346.

(41) Vandermause, J.; Torrisi, S. B.; Batzner, S.; Xie, Y.; Sun, L.; Kolpak, A. M.; Kozinsky, B. On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events. *npj Computational Materials* **2020**, *6*, 1–11.

(42) Sivaraman, G.; Gallington, L.; Krishnamoorthy, A. N.; Stan, M.; Csányi, G.; Vázquez-Mayagoitia, Á.; Benmore, C. J. Experimentally Driven Automated Machine-Learned Interatomic Potential for a Refractory Oxide. *Physical Review Letters* **2021**, *126*, 156002.

(43) Podryabinkin, E. V.; Shapeev, A. V. Active learning of linearly parametrized interatomic potentials. *Computational Materials Science* **2017**, *140*, 171–180.

(44) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less is more: Sam-

pling chemical space with active learning. *The Journal of chemical physics* **2018**, *148*, 241733.

(45) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chemical science* **2017**, *8*, 3192–3203.

(46) Huan, T. D.; Batra, R.; Chapman, J.; Krishnan, S.; Chen, L.; Ramprasad, R. A universal strategy for the creation of machine learning-based atomistic force fields. *NPJ Computational Materials* **2017**, *3*, 1–8.

(47) Sivaraman, G.; Guo, J.; Ward, L.; Hoyt, N.; Williamson, M.; Foster, I.; Benmore, C.; Jackson, N. Automated Development of Molten Salt Machine Learning Potentials: Application to LiCl. *The Journal of Physical Chemistry Letters 12*, 4278–4285.

(48) Ang, S. J.; Wang, W.; Schwalbe-Koda, D.; Axelrod, S.; Gómez-Bombarelli, R. Active learning accelerates ab initio molecular dynamics on reactive energy surfaces. *Chem* **2021**, *7*, 738–751.

(49) Warmuth, M. K.; Liao, J.; Rätsch, G.; Mathieson, M.; Putta, S.; Lemmen, C. Active Learning with Support Vector Machines in the Drug Discovery Process. *Journal of Chemical Information and Computer Sciences* **2003**, *43*, 667–673, PMID: 12653536.

(50) Freytag, A.; Rodner, E.; Denzler, J. Selecting influential examples: Active learning with expected model output changes. European Conference on Computer Vision. 2014; pp 562–577.

(51) Käding, C.; Rodner, E.; Freytag, A.; Mothes, O.; Barz, B.; Denzler, J.; AG, C. Z. Active Learning for Regression Tasks with Expected Model Output Changes. BMVC. 2018; p 103.

(52) Park, S.; Kim, T.; Yoon, S.; Koh, C. W.; Woo, H. Y.; Son, H. J. Progress in Materials, Solution Processes, and Long-Term Stability for Large-Area Organic Photovoltaics. *Advanced Materials* **2020**, *32*, 2002217.

(53) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society* **1996**, *118*, 11225–11236.

(54) Rühle, V.; Lukyanov, A.; May, F.; Schrader, M.; Vehoff, T.; Kirkpatrick, J.; Baumeier, B.; Andrienko, D. Microscopic Simulations of Charge Transport in Disordered Organic Semiconductors. *Journal of Chemical Theory and Computation* **2011**, *7*, 3335–3345, PMID: 22076120.

(55) Mckay, M. D.; Beckman, R. J.; Conover, W. J. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output From a Computer Code. *Technometrics* **2000**, *42*, 55–61, Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/00401706.2000.10485979.

(56) Plimpton, S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *Journal of Computational Physics* **1995**, *117*, 1–19.

(57) Neese, F.; Wennmohs, F.; Becker, U.; Riplinger, C. The ORCA quantum chemistry program package. *The Journal of Chemical Physics* **2020**, *152*, 224108.

(58) Webb, M. A.; Delannoy, J.-Y.; de Pablo, J. J. Graph-Based Approach to Systematic Molecular Coarse-Graining. *Journal of Chemical Theory and Computation* **2019**, *15*, 1199–1208.

(59) Horn, B. K. P. Relative orientation. *Int. J. Comput. Vis* **1990**, *4*, 59–78.

(60) Wilson, A. G.; Nickisch, H. Kernel Interpolation for Scalable Structured Gaussian Processes (KISS-GP). *CoRR* **2015**, *abs/1503.01057*.

(61) Gardner, J. R.; Pleiss, G.; Bindel, D.; Weinberger, K. Q.; Wilson, A. G. GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration. Advances in Neural Information Processing Systems. 2018.

(62) Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**,

(63) Kapoor, A.; Grauman, K.; Urtasun, R.; Darrell, T. Gaussian processes for object categorization. *International journal of computer vision* **2010**, *88*, 169–188.

(64) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703* **2019**,

(65) McKay, M. D.; Beckman, R. J.; Conover, W. J. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* **2000**, *42*, 55–61.

(66) Stukowski, A. Visualization and analysis of atomistic simulation data with OVITO–the Open Visualization Tool. *Modelling and Simulation in Materials Science and Engineering* **2009**, *18*, 015012.